R4.5 PDF Server Implementation Specification [Authors: Ellis, Mills, Yu]

Requirements

- Convert submitted materials with the following extensions to PDF/a.
  .doc (MS Word Document)
  .xls (MS Excel Spreadsheet)
  .ppt (MS PowerPoint Presentation)
  .pdf (PDF, non PDF/a)
  .tif/.tiff (Images)
  .html
- Validate that a submitted PDF complies with the PDF/a format or not.
- Merging of multiple PDF documents into a singular PDF.
- Integrate conversion of documents with appropriate workflow conditions in Faculty Deposit and ETD applications.

Implementation Scenarios

There are multiple implementation scenarios that need to be evaluated and eventually decided on. The following are the major steps involved in converting a document to PDF/a.

1) Delivery of document to be converted to PDF conversion server
2) Define conversion settings
3) Running and managing the PDF server
4) Return of converted document in PDF/a format to requesting application
5) File cleanup after successful conversion

- Delivery of document to be converted to PDF conversion server scenarios
  1) The PDF server allows for conversion of documents that are stored in a "watched" folder configured on the PDF server. Files would need to be delivered through a FTP/UNC mechanism to the PDF server for conversion.
  2) With XML Job Tickets/Job Info we can use Exponent calls AddJobFilesAsAttachments() & AddJobFilesAsStream() to stream files as either a DIME attachment (AddJobFilesAsAttachments) or an XML document (AddJobFilesAsStream).
  3) With XML Job Tickets/Job Info files can be referenced that live at an external course that is HTTP or FTP accessible. Exponent will download the file for conversion. File location should be specified in the Job Info file, not the XML Job Ticket in this case. This will allow for more efficient queuing of documents for conversion.

- Define conversion settings
  Using XML Job tickets conversion settings can be defined when the document is delivered for processing. Additionally XML Templates can be referenced in a XML Job Ticket. The XML Template would store partial Job Ticket settings that might apply to a majority of processing requests.

- Running and managing the PDF server
  There are two PDF servers, development and production. Appropriate licenses have been purchased for both servers. The development and production both share identical

specifications to aid in simulating response times given certain loads during testing. All development and testing work will be done against the development server and all production work will be directed at the production PDF server. The development PDF server will be housed at the SCC office and the production server will be housed at the Systems office. Sho Nakagama will administrate the SCC server. Anne Butman and Nick Gonzaga will serve as administrators for the production PDF server. They will maintain communication amongst each other to provide both environments stay as similarly configured as possible. The development server is currently configured and installed. The production server will be installed and configures and ready for the R4.5 release.

- Return of converted document in PDF/a format to requesting application
  1) The generated PDF/a can be outputted to a web accessible address on the PDF server that the requesting application could then get the file from. This would require the PDF server to have a web accessible workspace available.
  2) The XML Ticket could define FTP/UNC information (path/username/password) the generated PDF could be sent to after generation is complete by the PDF server software.

- File cleanup after successful conversion
  The PDF server is not intended as a permanent storage location for generated PDF's. Once a document has been converted and submitted and returned successfully to the requesting application some file cleanup will need to be performed. Using setting in either the Express "watched" folder setting or invoking DeleteJobFolder() the source document that was converted could be deleted. What will be left is the generated PDF on the PDF server in a specified "output" location. Since this PDF should have already been delivered to the requesting application it is no longer needed. There is no feature in the PDF server software to accomplish this. Two options are possible.
  1) Requesting application cleans up this file as well through a designed solution.
  2) An automated policy is implemented on the PDF server that will remove files considered to be expired, no longer needed.

Additional items

Metadata stored in the PDF

Using XML Job ticketing the ability to store some metadata in the generated PDF exists. The fields available are:

- Title
- Subject
- Author
- Keywords
- User Defined

If these field(s) were to be used a mapping would be needed from the submitted metadata to these metadata fields stored in the PDF document.

XML Job Ticket Tool

XML job tickets are used during a conversion of a document. It complies to a DTD supplied by Adlib. It would behoove us to create a tool that can read in the DTD

supplied and generate a GUI for XML Job Ticket creation. This tool would be intended for use by developer's and not the average end user.

PHP module for using PDF server

To reduce the reinventing of the wheel a unified class of methods for submission of documents to be converted/validated/merged from PHP apps with SOAP protocol etc. needs to be written and shared amongst applications using the PDF server.

OCR

In some instances TIFFs or PDFs without a text layer (e.g. scanned documents) may need to be processed by OCR software to make text extraction possible. Although there is an existing OCR workflow used by WMS, RUetd would benefit from the use of Adlib's built-in OCR tools. The need for OCR processing would be determined using document metadata.